图书情报工作
第 63 卷 第 7 期 2019 年 4 月

基于特征分解的知识网络结构关系提取*

栾宇 滕广青 安宁 韩尚轩
东北师范大学信息科学与技术学院 长春 130117

摘要:[目的/意义]对知识网络中结构关系的有效识别与提取,有助于从纷繁的数据中探测知识网络的拓扑结构及其演化模式。[方法/过程]本文提出一种基于邻接矩阵特征分解的知识网络结构关系提取方法。基于真实数据分别从静态结构关系提取和动态结构演化两个方面,对特征分解法和传统关联频度法进行对比分析,并与 Pathfinder 算法进行对比。对基于特征分解法提取知识网络结构关系的有效性进行验证。[结果/结论]研究结果表明:特征分解法能够识别原始知识网络中的主要成分信息,能够准确识别低频次的对网络整体拓扑结构较为重要的关联关系,且提取方法灵活自由。

关键词:知识网络 特征网络 特征分解 结构关系
分类号:G255.76
DOI:10.13266/j.issn.0252-3116.2019.07.012

1 引言

知识网络分析方法已经成为图书情报领域的一种探究知识单元之间关联关系与结构模式的新型研究范式。研究人员以关键词、专利技术等知识单元作为网络节点,以知识单元之间的关联关系作为边,构建关键词网络、专利网络等知识网络。通过分析知识网络的统计特征及拓扑结构,可以研究知识的层级结构与演化机制等问题。随着大数据分析思想在科学研究中的应用,知识网络研究中的数据规模庞大与多样性等特征纷纷得以体现。但是在知识网络研究关注数据分析的高价值的同时,不得不面对大数据分析中的低价值密度问题。因此,在保障数据信息高频显著性的同时,提取知识网络的有效结构关系,成为知识网络分析中一个亟需解决的问题。

针对传统关联频度等方法在提取网络信息方面的不足,本文提出一种兼顾网络关联频度与拓扑信息的知识网络结构关系提取方法,并分别通过静态与动态的知识网络结构关系提取,对该方法的有效性进行验证。

2 相关研究工作

随着 20 世纪末网络科学(Network Science)[1]几个重要研究成果的发表,网络分析作为一种新的研究范式越来越受到学术界的重视。网络思维尤其适用于针对复杂系统中的组成元素及其关联关系进行描述和分析,即将系统组成元素抽象为网络节点,将元素之间的关联关系抽象为网络的边。网络作为对这些复杂系统的一般抽象和描述方式,突出强调了复杂系统的拓扑结构与统计特征[2]。网络科学与图书情报学的结合也改变了图书情报学传统的基于描述统计的研究范式。研究人员在专利分析[3]、引文分析[4]、热点判识[5]、知识涌现[6]、重要学者识别[7]、知识网络结构[8]等领域引入网络分析理论,更好地揭示了研究对象的结构关系。J. Liebowitz[9]指出,网络分析手段有助于更好地解释知识之间的关联关系和知识的层级结构,在知识管理中的作用日益突出。

在研究工作以数据为基础进行分析和挖掘的同时,随着数量的增加,研究工作必须面对大数据分析中价值密度低的问题。早期的节点频次法是首先选取领域中高频出现的知识单元(如关键词)作为网络节点,

* 本文系国家自然科学基金面上项目"基于网络结构演化的 Folksonomy 模式中社群知识组织与知识涌现研究"(项目编号:71473035)研究成果之一。
作者简介:栾宇(ORCID:0000-0003-3752-2470),硕士研究生;滕广青(ORCID:0000-0002-1053-0959),教授,博士生导师,通讯作者,E-mail:tengguangqing@163.com;安宁(ORCID:0000-0002-9579-0150),硕士研究生;韩尚轩(ORCID:0000-0001-0962-3218),博士研究生。
收稿日期:2018-07-10 修回日期:2018-11-08 本文起止页码:96-104 本文责任编辑:王传清

然后建立节点之间的关联关系[10]。该方法强调知识单元(节点)个体的高频性,但是对知识关联关系的重视程度不足。目前采用较多的方法是关联频度法,首先选取高频次的关联关系(节点对),基于高频关联关系提取建立知识网络。满足特定关联频度的知识网络相比节点频次法凸显了网络分析中关联关系的重要性。这种提取网络信息的方法简单直观且计算任务少,被广泛应用于图书情报研究领域,在关联主题提取[11]、知识网络结构识别[12]等相关研究中取得了相应的成果。学术界将这些基于一定关联频度水平提取的知识网络称为水平网络或层次网络(Network at Level)[13]。

在对网络进行简化方面,除了基于节点关联频度的提取方法外,还有一类以 Pathfinder 算法[14]为代表的基于网络路径的提取方法。Pathfinder 算法中,如果两个节点之间的距离 D 小于直接连接两个节点的边的权重,则将这条边从网络剔除。因此 Pathfinder 算法的主要思想是通过计算节点之间不同路径的权重,来比较判断节点之间直接相连的边的重要程度。该算法在知识网络路径研究和简化分析中最典型的应用是文献分析软件 CiteSpace。陈超美[15]将 Pathfinder 算法引入到知识网络分析中,借助 Pathfinder 算法的完备性,在简化网络的同时突出重要的链路特征。由于 Pathfinder 算法在主路径识别方面的独特性[16],该方法在社区知识结构[17]、成员结构[18]、知识交流路径[19]、关键文献识别[20]等方面也被广泛应用。

随着网络科学研究在网络拓扑结构方面成果的发布与积累,上述曾经得到学术界认可的关联频度法和 Pathfinder 算法等,在突出关联关系频度或链路重要性的同时,也在一些针对特定问题的网络拓扑结构信息的提取方面显露出不足之处,可能会使研究工作遗漏网络的重要拓扑结构信息。鉴于此,本文提出一种基于邻接矩阵特征分解的知识网络结构关系提取方法,以期在生成的知识网络中兼顾关联频度与重要拓扑结构信息。

3 理论与方法

3.1 知识网络结构关系的相关理论

知识单元一般被认为是领域知识的最小单位,但是单一知识单元的属性并不能表征领域知识的整体属性特征。正如《信息简史》的作者 J. Gleick[21]所指出的,知识之间的连通性比知识单元本身更为重要。领域内诸多知识单元个体基于一定的关联关系形成知识网络宏观上的结构模式,这恰恰是针对复杂系统研究所遵循的主要视角。知识网络研究正是基于网络思维对知识单元尤其是知识关联关系进行考查。大量的知识网络研究工作表明,众多的知识单元通过特定的关联关系聚集在一起,形成特定的拓扑结构特征,进而涌现出知识网络宏观层面的模式与规律。

提取知识网络结构关系的传统方法主要是节点频次法、关联频度法和 Pathfinder 算法。节点频次法将单一知识单元(网络节点)的重要性放在了首位。李纲等[22]指出,用高频词代表领域整体的研究方向存在着天然的缺陷,而低频词有助于获取一些隐含主题或前瞻主题的信息。而关联频度法在获得关联频度阈值水平(层次)上的知识网络过程中,对关联频度的考查仅限于关系(节点对)个体,没能兼顾整体网络的拓扑结构,也可能在一定程度上忽视或遗漏重要的低频度知识关联。Pathfinder 算法只能在节点之间存在路径权重较低的路径的前提下剔除节点之间的连边,因此 Pathfinder 算法的计算结果是一个连通的网络,原始网络中的节点规模并不会得到精简。在知识网络研究中,知识节点(如关键词)之间的关联频度只是知识网络的局部信息。判断知识节点及其关联关系的重要程度,不仅要参考知识节点及其关联关系的局部信息,还要从网络整体拓扑结构上对其加以考量。在领域知识演化发展的过程中,即使低频度的关联关系(节点对)有时也扮演着非常重要的角色。

邻接矩阵是知识网络的数学表达。对邻接矩阵进行特征分解,能够在关注知识单元与知识关联的局部属性同时,兼顾知识网络拓扑结构的整体属性。因此,本文提出基于邻接矩阵特征分解的知识网络结构关系信息提取方法。特征分解(eigen decomposition)指的是将一个矩阵分解为由其特征值和特征向量表示的矩阵之积。原始知识网络 G(N,L)的邻接矩阵 A 是一个实对称方阵,因此能够对其实施特征分解。矩阵可以被理解为在其线性空间下的一种线性变换的描述。通过特征分解,能够发现矩阵的特征值及特征向量,由此推导得出矩阵所描述的变换形式,并提取矩阵的主要特征值所对应的结构关系信息。邻接矩阵 A 承载了原始知识网络 G 的全部连接信息,其中元素的值是知识节点之间的关联频次,特征值在网络中有对应的拓扑结构信息。提取邻接矩阵 A 的主要特征值,相当于融入 PageRank[23]算法思想后提取网络 G 的拓扑结构信息。据此可以组合生成研究工作需要的特征网络。

3.2 方法与流程

基于真实数据构建的邻接矩阵 A 是一个多值矩阵。矩阵中的元素值表示知识节点之间的关联频次。对于邻接矩阵 A，如果元素数值为 0，则相对应的知识节点之间不存在关联关系；如果元素值大于 0，则相对应的知识节点之间存在关联关系。对于一个 N＊N 的邻接矩阵 A，它有 N 个实特征值，降序排列记作 λ_i（i＝1，2，3，…，N），同时对应有 N 个线性无关的特征向量，记作 q_i（i＝1，2，3，…，N）。对邻接矩阵 A 进行特征分解，得到：

$$A = Q\Lambda Q^T$$

其中，Q 是由特征向量组成的正交矩阵，Q 的第 i 列等于 q_i^T，Λ 是对角线元素为降序排列的特征值 λ_i（i＝1，2，3，…，N）的对角矩阵。

与此同时，对邻接矩阵 A 进行二值化处理，得到二值矩阵 C，用于判断知识节点间关联关系的存在性。根据研究需要，选取若干能够代表邻接矩阵 A 的主要变换形式的特征值，将对角矩阵 Λ 中除需要保留特征值之外的其他特征值赋值为 0 得到 Λ_B，由此构建矩阵 B：

$$B = Q\Lambda_B Q^T$$

矩阵 B 是对邻接矩阵 A 的信息的提取，其中包含有邻接矩阵 A 的特定的特征值及其特征向量所对应的变换信息。矩阵 B 中的元素的值是相对于特定的特征值组合下网络边的权重。影响矩阵 B 中元素值的因素有两个：一是知识节点之间的关联频次；二是特定的特征值所对应的网络拓扑结构信息。进一步将矩阵 B 与二值矩阵 C 进行对照，如果二值矩阵 C 中某关联关系不存在，则对矩阵 B 进行修正，将对应的边权重修正为 0。修正后的矩阵 B 对应的知识网络的拓扑结构与原始知识网络一致，只是此时知识网络的边权重被重新赋值。与传统关联频度法中的固定权重不同，特征分解法中边的权重是一种动态权重，具有相对性。选定的特征值不同，所提取的拓扑结构信息不同，网络中每一条边的权重的值都会相对产生变化。这样定义的边的权重兼顾了网络的局部信息和整体拓扑结构信息。

接下来，根据边的权重设定阈值，舍弃重新赋值的知识网络中权重低于阈值的边，并剔除孤立节点，生成特征网络 E，至此完成对知识网络结构关系信息的提取。整个流程见图 1。

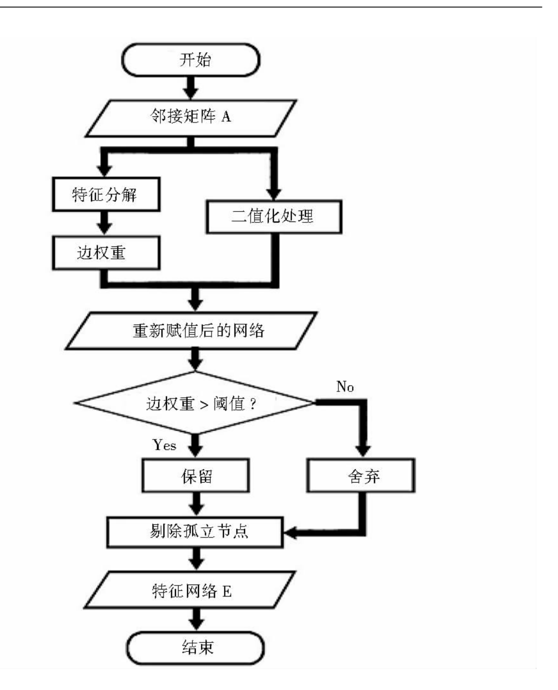


图 1 特征分解法提取网络结构关系流程

需要注意的是，基于真实数据而构建的邻接矩阵 A 往往比较稀疏，这使得邻接矩阵 A 中存在少量数值较大的特征值和大量数值较小的特征值。因此只需少量的特征值及特征向量组合即可描述邻接矩阵 A 的主要变换内容。对于原始知识网络 G 而言，这些少量的特征值及特征向量对应了网络的重要拓扑结构。特征值的选取决定了网络中哪些拓扑结构信息将被提取出来，可以选取任意序位的一个或多个连续的或不连续的特征值。此外，将矩阵 B 中的元素值作为边的权重赋值给原始知识网络 G 的边集合 L，而边权重的阈值用于判定和提取相对于已选定的拓扑结构而言重要程度较高的边。特征值选取决定了网络中被提取的拓扑结构信息，边权重阈值则决定了最终提取得到的特征网络的规模。如果需要提取较多的网络信息，可以增加特征值数量，适当降低权重阈值；反之则可以减少特征值数量，适当升高权重阈值。在特征分解法中，被选定的特征值组合所对应的网络拓扑结构中的边的权重得到提升，其他边的权重则相对降低，从而能够根据不同的关注重点突出网络局部结构的重要程度。正因为这一特性，特征分解法不仅可以从整体上对网络进行简化，还能够识别、提取网络中的重要局部信息。

3.3 特征分解法与关联频度法

在以往利用关联频度法的研究中，阈值的设定往往需要依赖于研究者的个人经验，阈值的选取局限于正整数域。反映在对邻接矩阵 A 的操作上，则是将矩阵 A 中低于阈值的元素重新赋值为 0。深入对比分析特征分解法与关联频度法提取网络信息的过程，可以发现关联频度法实际上是特征分解法的一个特例。在特征分解法中，当矩阵 A 的全部特征值都被保留时，

Λ_B 等于 Λ,此时矩阵 B 与矩阵 A 相等。如果将边的权重阈值设为与关联频度阈值相同的正整数,则特征分解法的返回结果与关联频度法相同,此时两种提取方法是在以一个相同阈值对同一矩阵 A 进行操作。

很明显,关联频度法就是选取邻接矩阵的全部特征值,并将边的权重局限于正整数域的特征分解法。从特征分解的角度看,关联频度法基于网络的全局视角,选取邻接矩阵的全部特征值,以正整数阈值对网络进行简化,是科学合理的。但是,关联频度法作为特征分解法的一个特例,忽视了真实世界中网络的邻接矩阵的稀疏性问题,无差别地选取全部特征值,是无法识别、提取特定的局部网络信息成分的。同时,关联频度法的阈值被天然地局限于正整数域。关联频度法只能在阈值 2 的水平上提取最多的网络信息,这显然不利于研究人员对知识网络进行深入分析。特征分解法则是关联频度法的一般性扩展。特征分解法突破了关联频度法在特征值和阈值选取上的局限性,能够更为灵活、准确地提取网络信息。

4 实例分析

4.1 研究数据

研究中的数据采集自中国知网(CNKI)和万方数据(WANFANG DATA)知识服务平台。以“知识管理”为检索主题,以图书情报为学科领域,检索 1999-2017 年间发表于 CSSCI 期刊的全部论文,提取论文题目、关键词、发表时间等相关信息。为了从不同知识群落之间关联关系的角度更好地对知识结构关系提取方法进行比较分析,进一步以“社会网络”“复杂网络”“社会网络分析”“网络分析”和“网络科学”为检索关键词,在中国知网中以“自然科学理论与方法”“社会科学理论与方法”“数学”和“非线性科学与系统科学”为学科领域,在万方数据中以“自然科学总论”“社会科学总论”和“数理科学和化学”为学科领域,检索 1999-2017 年间发表于 CSSCI 期刊的全部论文,提取论文题目、关键词和发表时间等信息。汇总上述检索所采集的全部数据,剔除重复论文、会议通知、期刊公告等无效数据,最终得到期刊论文合计 1 842 篇,关键词合计 3 018 个。从 1999-2001 年度起,以 3 年作为一个时间窗口,以 1 年作为步长,平滑移动至 2015-2017 年度,得到 17 个时间窗口的相关基础统计数据,见表 1。

考虑到领域知识网络的时间序列分析中,当期发生值与累计值所反映的侧重点之间的差异性,以及本研究对知识关联关系生长与消退的重点关注,表 1 中的数据为各个时间窗口(3 年)的当期发生值。同时,为了避免数据剧烈波动对领域知识发展的影响,在时间序列上移动一步(1 年)进行“修匀”。通过多个检索主题与关键词的跨平台组合(CNKI 和 WANFANG DATA),以及移动平滑的时间窗口划分,可以在突出知识单元与知识关联固有发展规律的同时,显示出“知识管理”知识群落与“社会网络”“复杂网络”等知识群落在时间序列下的交叉聚合等演化现象。

表 1 各时间窗口下论文及关键词数量

时间窗口	论文篇数	关键词个数	时间窗口	论文篇数	关键词个数
t1	56	126	t10	419	865
t2	112	202	t11	400	852
t3	162	267	t12	395	839
t4	227	376	t13	388	899
t5	273	498	t14	371	917
t6	319	617	t15	342	919
t7	367	725	t16	313	838
t8	410	822	t17	267	748
t9	436	873			

4.2 提取流程

首先,根据表 1 中的相关数据构建原始邻接矩阵与原始网络。分别统计 t1-t17 时间窗口下的关键词在文献中的共现关系及频次,并据此建立各个时间窗口的关键词多值邻接矩阵 A_{t1}-A_{t17}。根据 17 个邻接矩阵,基于关键词共现关系分别生成 17 个时间窗口的领域知识网络 G_{t1}-G_{t17}。其次,根据常规的关联频度法提取知识网络的结构关系信息。在关联频度法中,阈值越小,提取出的知识网络结构关系信息就越细腻。研究中,以 2 作为频度阈值,过滤掉原始网络中共现频次小于 2 的关联关系,即剔除邻接矩阵 A_{t1}-A_{t17} 中小于 2 的元素,由此获得相应频度水平的知识网络 F_{t1}-F_{t17}。再次,根据特征分解法提取原始网络的结构关系信息。对邻接矩阵 A_{t1}-A_{t17} 进行特征分解,以不同组合方式提取矩阵特征值,根据设定的阈值提取知识网络的结构关系信息,生成相应特征值下的特征网络 E_{t1}-E_{t17}。最后,将基于特征分解提取的特征网络(E_{t1}-E_{t17})与原始网络(G_{t1}-G_{t17})和关联频度法提取的水平网络(F_{t1}-F_{t17})进行对比分析,以验证特征分解法在识别、提取网络主要结构关系信息中的功能与特性。

以 t17 时间窗口下的知识网络 G_{t17} 为例。对其邻接矩阵 A_{t17} 进行特征分解。邻接矩阵 A_{t17} 共有 748 个特征值。将邻接矩阵 A_{t17} 的特征值由大到小排序编号,以顺序号为横轴,特征值为纵轴,矩阵 A_{t17} 的 748 个

特征值排列如图 2 所示：

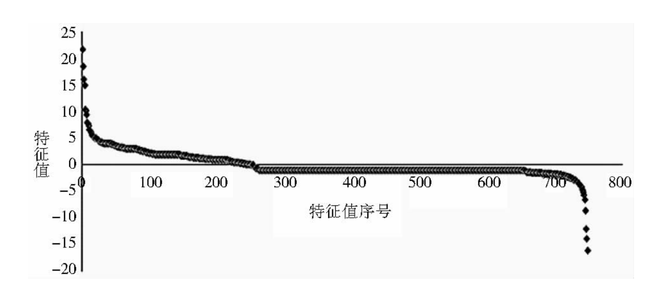


图 2　邻接矩阵 At17 的特征值

从图 2 可以观察到，大多数的矩阵特征值集中在横轴（0 值）附近，只有少数的特征值偏离横轴较远（绝对值较大）。由矩阵特征分解的原理可知，少量的绝对值较大的特征值对应了网络 G_{t17} 中的主要拓扑结构信息。因此在基于网络思维的知识结构关系的提取中，只需提取这些重要程度较高的特征值（绝对值较大的特征值），就可以涵盖领域知识之间主要的结构关系。

知识结构关系的提取过程中，重点提取绝对值较大（图 2 中偏离横轴较远）的序位靠近极值的特征值，以 0.5 作为边的权重阈值，提取原始知识网络的结构信息，基于所提取的结构信息组合生成基于特征分解的特征知识网络。提取邻接矩阵 A_{t17} 的最大、次大特征值（序号为 1、2）和最小、次小特征值（序号为 747、748），组合生成的特征网络如图 3 所示：



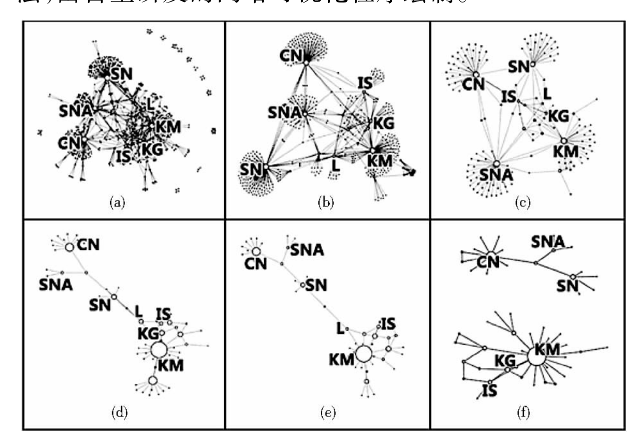
图 3　部分特征值组合生成的特征网络

结合图 3 中的结果和矩阵 A_{t17} 中的信息可以发现，原始知识网络 G_{t17} 的主要特征值（序号为 1、2、747、748）对应的网络拓扑结构信息为知识管理（KM）和社会网络（SN）子网，两个知识群落之间的结构关系在图 3 中得到清晰地呈现。经验证，单独提取某一特征值所得到的子网络中的节点与边均来自于原始知识网络 G_{t17} 的节点集合 N_{t17} 和边集合 L_{t17}。这与数据采集内容和采集方案是相符合的，说明特征分解法能够识别知识网络中的主要结构关系信息。由此，可以根据研究需要决定所要提取的特征值组合。

5　分析结果

5.1　静态网络结构关系提取

在特征分解法中，确定了需要提取的特征值组合，则确定了被提取出来的相关网络的结构关系信息。而边的权重阈值则决定了最终从原始网络中提取出的子网的规模。以 t17 时间窗口为例，对 A_{t17} 进行特征分解，提取第 1 – 11 和第 738 – 748 共 22 个绝对值较大的特征值，边的权重阈值分别设定为 0.5、1.0、1.1 和 1.2。基于这 22 个特征值的网络结构关系信息，分别组合构建不同边权重阈值下的特征网络，并与原始知识网络 G_{t17} 和关联频度法提取的水平网络 F_{t17} 对比，所得结果如图 4 所示。图 4 中的网络采用 T. M. J. Fruchterman 和 E. M. Reingold[24] 提出的 F-R 布局算法，由自主研发的网络可视化程序绘制。



注：KM – 知识管理，SN – 社会网络，CN – 复杂网络，SNA – 社会网络分析，KG – 知识图谱，IS – 情报学，L – 图书馆

图 4　原始网络与特征网络和水平网络

图 4 中，（a）为原始知识网络 G_{t17}，（b）为阈值 = 0.5的特征网络，（c）为阈值 = 1.0 的特征网络，（d）为阈值 = 1.1 的特征网络，（e）为阈值 = 1.2 的特征网络，（f）为水平知识网络 F_{t17}。同时，基于特征分解法，生成特征网络 E_{t17} 与原始网络 G_{t17} 和水平网络 F_{t17} 的基本统计特征，如表 2 所示：

表 2　网络统计特征

网络		点数	边数	密度	聚类系数	特征路径长度
原始网络 G_{t17}		748	2 054	0.0074	0.1664	3.3612
特征网络 E_{t17}	阈值 = 0.5	591	930	0.0053	0.0743	3.1906
	阈值 = 1.0	147	183	0.0171	0.0262	3.1083
	阈值 = 1.1	56	68	0.0442	0.0800	3.8617
	阈值 = 1.2	51	60	0.0471	0.0750	4.3388
水平网络 F_{t17}		51	57	0.0447	0.0610	2.6267

结合图 4 和表 2 中的相关数据可以发现，在特征

分解法中，确定提取邻接矩阵的特征值后，边的权重阈值越低，特征网络就越接近原始网络 G_{t17}；权重阈值越高，特征网络在保留原始网络主要结构特征的情况下就越精简。在本实例中，当边的权重阈值取 1.2 时，所获得的特征网络的节点数和边数近似于关联频度阈值＝2 时的水平网络 F_{t17}。然而图 4 中(e)为连通网络，(f)为非连通网络。显然，特征网络更倾向于保留呈现网络的主要结构关系，而水平网络则侧重于网络节点间关联频度的统计显著性。特征网络的这一特点在领域知识发展演化的动态网络分析中，将有助于知识关联关系生长或衰退的呈现与揭示。

此外，特征分解法中的边权重阈值既可以取整数，也可以取小数，在取值的选择上要远远丰富于关联频度法。关联频度法中，频度阈值可选定的最小值为整数 2，说明最小只能以阈值＝2 的水平从原始网络中提取信息(频度阈值越小从原始网络中提取到的信息越丰富)。特征分解法通过灵活组合设定特征值和阈值，不局限于固定阈值，能够自由提取、缩放原始网络中的整体或局部拓扑结构特征。

5.2 结构关系动态演化揭示

通过提取绝对值较大的特征值，能够识别出原始知识网络中的主要结构关系信息。由知识网络的静态分析可以发现，相同边权重条件下，知识管理(KM)、社会网络(SN)和复杂网络(CN)等子网的规模相对较大，与数据采集方案相符合。另一方面，知识关联关系的涌现能够影响知识网络拓扑结构的演化[25]。因此，进一步采用特征分解法提取 1999－2017 年间原始知识网络(G_{t1}-G_{t17})中知识管理、社会网络和复杂网络对应的特征值，组合生成时间序列特征网络(E_{t1}-E_{t17})。重点对网络中 3 个知识群落的交叉关联状态进行跟踪，并与基于关联频度法得到的时间序列水平知识网络(F_{t1}-F_{t17})进行比较分析。

以特征分解法来提取 17 个时间窗口下的原始网络的结构关系信息。首先，从 17 个时间窗口的邻接矩阵中提取与知识管理、社会网络和复杂网络相关的特征值。然后，以 0.5 作为权重阈值，在 17 个时间窗口下分别提取原始知识网络的相关特征值组合生成相应的特征网络。同样采用 F-R 布局算法[24]，阈值＝0.5 的时间序列特征网络如图 5 所示。

图 5 中的时间序列特征网络显示，在时间轴的初始阶段(t1、t2)，考察对象中的社会网络(SN)与知识管理(KM)知识群落之间尚未连通，处于各自发展阶段。此时复杂网络(CN)相关知识尚未在考察对象中出现。

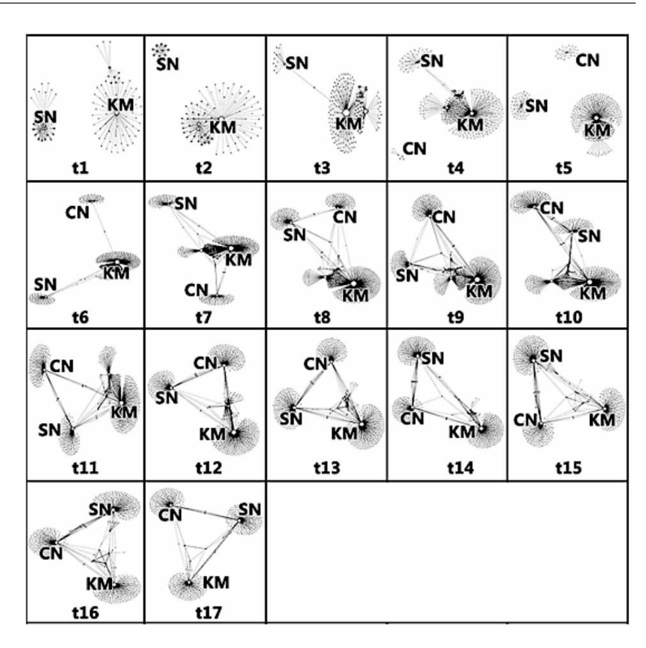


图 5　阈值＝0.5 的时间序列特征网络

至 t3 时间窗口，知识管理与社会网络知识群落之间首次连通。t4 时间窗口中，复杂网络知识群落首次出现，但与知识管理和社会网络知识群落之间处于非连通状态。t5 时间窗口中，上一时间窗口处于连通状态的知识管理与社会网络知识群落之间再次断开，此时 3 个知识群落之间互不连通。t6 时间窗口中，社会网络和复杂网络知识群落分别与知识管理知识群落连通，即社会网络与复杂网络只有通过知识管理才能建立联系。从 t7 时间窗口开始，社会网络、复杂网络、知识管理 3 个知识群落通过多个知识节点或其他小规模群落始终保持相互连通状态。在后续的时间窗口中，社会网络、复杂网络、知识管理 3 个知识节点较稳定地形成了网络中由桥点(bridge)相互连通(或直接关联)的峰点(peak)[26]，三者之间的关联关系得到清晰地呈现。

为了对特征分解法在不同边权重下的提取效果有更清晰地认识，研究中进一步以 1.0 为边权重阈值，生成特征网络，见图 6。

图 6 中所提取的特征值与图 5 相同，不同之处在于边权重阈值取值为 1.0，据此组合生成时间序列特征网络。将图 6 与图 5 对比分析可以发现，由于边权重的提高，一些低于阈值的边被舍弃，相应产生的孤立节点被排除。由此导致 3 个知识群落的规模相应地减小，一些在低权重阈值下可见的结构关系也相应地隐藏，如 t3、t4、t6、t7、t8、t9 等时间窗口。尽管如此，图 6 对知识管理、社会网络和复杂网络知识群落之间结构关系的揭示总体上与图 5 一致，在时间序列的后期形成较为鲜明的由桥点相互连通(或直接关联)的 3 个峰点。

图 6　阈值 = 1.0 的时间序列特征网络

出于对比验证的目的，进一步根据一般的关联频度法从原始知识网络 G_{t1}-G_{t17} 中提取网络中节点关系信息。以 2.0 作为关联频度阈值，从原始网络中提取关联关系及相应的节点对，并生成关联频度在 2.0 这一水平上的水平网络 F_{t1}-F_{t17}，如图 7 所示：

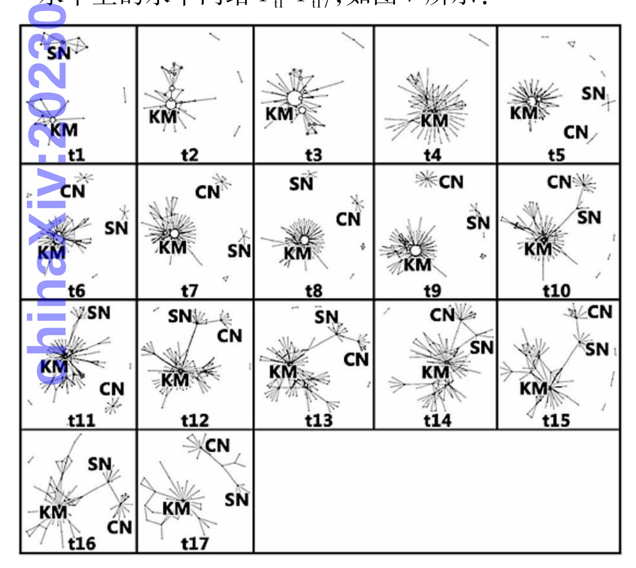


图 7　阈值 = 2.0 的时间序列水平网络

关联频度法中的关联频度阈值为 2.0 意味着除了原始网络（相当于关联频度阈值为 1.0）之外最小的阈值，所生成的水平网络最接近原始网络。由此生成的水平网络（见图 7）的分支结构（非连通碎片）明显多于特征分解法生成的特征网络（见图 5、图 6）。在图 7 的所有时间窗口中，关联频度法生成的水平网络都存在非连通的碎片，说明一些重要的结构关系被关联频度阈值拒绝。时间轴的前半段（t1-t9），知识管理、社会网络和复杂网络 3 个知识群落之间始终未能建立联系，处于非连通状态。在时间轴的后半段，t10、t12、t13、t15、t16 时间窗口中复杂网络只能通过社会网络才能与知识管理知识群落建立联系；t11 时间窗口知识管理与社会网络连通，复杂网络与两者非连通；t17 时间窗口社会网络与复杂网络连通，知识管理与两者非连通；仅有 t14 时间窗口呈现出特征分解法中时间轴后半段表现显著的 3 个峰点通过桥点相互连通（或直接关联）的现象。

事实上，知识的交叉融合现象是知识发展过程中的普遍规律。自 20 世纪末网络科学复兴以来，网络分析作为一种新的研究范式已经被引入到知识管理研究的诸多方面，社会网络、复杂网络与知识管理在研究工作中的交叉结合已经见诸于众多文献。在基于特征分解法提取的结构关系组合生成的特征网络中，能够观察到社会网络、复杂网络和知识管理 3 个知识群落之间的交叉关联的具体形成过程。关联频度法在边的权重设定上只考虑节点之间的关联频度，没有考虑网络整体拓扑结构属性，不能识别频次低于阈值但在整体拓扑结构中处于相对较重要位置的边的信息（具体表现为图 7 中多个时间窗口的非连通及“峰点”间桥接关系丢失现象）。因此，在由满足频度阈值的边（节点对）生成的水平网络中，知识网络演化过程中的低频次的重要关联关系将会丢失。

5.3　特征分解法与 pathfinder 算法对比

在分析了特征分解法与关联频度法差异的基础上，进一步将特征分解法与另外一种基于路径的简化网络算法 Pathfinder 算法进行对比分析。首先，根据 Pathfinder 算法对 t2、t7、t12 和 t17 时间窗口下的知识网络进行简化。将参数 q 设定为 n-1，即考虑节点之间的所有真实存在的路径。分别将参数 r 设定为 1 和无穷大（Inf），观察在不同的参数 r 的作用下，Pathfinder 算法对网络的精简效果。

表 3　Pathfinder 子网（PFNETs）统计特征

网络类型		t2		t7		t12		t17	
		点数	边数	点数	边数	点数	边数	点数	边数
原始网络		202	546	725	2 090	839	2 482	748	2 054
PFNETs	r = 1	202	526	725	2 029	839	2 429	748	2 044
	r = Inf	202	500	725	1 938	839	2 323	748	1 997

Pathfinder 算法在简化网络的过程中会至少保留一条原本连接两个节点的路径，因此不会破坏网络原本的连通性，提取得到的子网络 PFNETs 的节点数量与原网络保持一致，节点规模没有精简。从边的数量上看，子网络 PFNETs 对于网络的简化程度也低于特

征分解法和关联频度法。Pathfinder 算法适合用来解决关于网络主要路径的一些问题。在网络拓扑结构信息的识别与提取上，特征分解法的灵活之处还在于可以通过选取特征值组合和设定阈值，自由地决定所要提取的网络成份及规模，这是 Pathfinder 算法无法实现的。

6 结论与讨论

本文提出一种基于邻接矩阵的特征分解提取知识网络结构关系的方法。研究中通过所提取的特征值组合生成特征网络，并分别从静态分析和动态分析两个方面，对特征分解法和关联频度法进行了比较分析。分析结果表明，基于特征分解法的知识网络结构关系提取具有如下特性。

（1）特征分解法能够识别知识网络中的主要成分信息。从知识网络提取结构关系信息，原则上要在考虑知识网络的整体拓扑结构性质的前提下，判断网络中的节点与边的重要程度，剔除重要性相对较低的边。以往根据节点出现的频次或节点对关联频度设定阈值提取网络结构关系信息的数据处理方法，是根据微观层面知识单元或知识关联的局部信息来判定宏观层面的领域知识的整体属性，其正确性和有效性均不能满足全局网络结构关系呈现的需要。这两种传统方法忽略了网络的拓扑结构性质，一定程度上偏离了网络科学思维的初衷，有可能导致研究者错失网络的重要结构信息，处于萌芽状态（低频次）的知识单元或知识群落之间的交叉关联往往被忽略。

（2）特征分解法能够准确识别低频次的但对网络整体拓扑结构较为重要的关联关系，对知识之间的结构关系比较敏感。在领域知识生长发展的进程中，新知识的加入和新关系的建立都需要经历由弱到强的过程。网络整体结构的改变往往来自于细小变化的反复叠加[1]，因此低频次的关联关系也可能对领域知识网络拓扑结构的改变起到至关重要的作用。特征分解法保留了拓扑层面上的重要路径，完好地呈现了社会网络、复杂网络和知识管理 3 个通过桥点相互连通（或直接关联）的峰点（见图 5、图 6）；关联频度法则明显拒绝了复杂网络与知识管理之间的低频度关联（见图 7）。显然，单纯对关联频度的关注会一定程度地降低网络拓扑层面关键细节的呈现，而特征分解法能够在所提取的网络中观察到指定特征值对应的节点之间交叉关联的细微过程。因此，特征分解法对于知识单元或知识群落之间的交叉关联更为敏感，对研究知识之间的动态演化机制更有帮助。

（3）特征分解法能够灵活自由地提取知识网络结构关系信息。关联频度法以知识单元之间关联的频次作为判定知识单元之间的关联关系的重要程度的依据。特征分解法在知识单元关联关系的权重与阈值的设定问题上，综合考虑了网络的整体拓扑结构性质和局部的节点之间的关联关系，较关联频度法更为优秀。此外，特征分解法通过矩阵特征值的选择，既能够借助绝对值较高的特征值生成反映原始网络拓扑结构的特征网络，也能够针对研究重点选取指定特征值生成研究者需要的特征网络。这一特点使得基于特征分解法提取知识网络结构关系信息具备灵活自由的特点，既可以用于对知识网络整体拓扑结构的判识，也可以对特定知识关联的细微变化进行研判。

虽然特征分解法能够准确识别和提取知识网络中特定的信息，但相对于传统的网络数据处理方法，其实现过程较为复杂。在后续的研究中，将继续思考特征分解法在知识网络研究领域的应用模式，进一步规范、简化特征分解法的具体步骤。同时，在对邻接矩阵进行特征分解的过程中初步发现，相对于矩阵的主要特征值，绝对值较小的特征值所对应的网络拓扑结构也在一定程度上影响着知识网络的统计特征。对于这些绝对值较小的特征值在知识网络中的指代意义的探索也是接下来的研究工作。

参考文献：

[1] LEWIS T G. 网络科学：原理与应用[M]. 陈向阳，巨修练，等译. 北京：机械工业出版社，2011.
[2] 郑金连，狄增如. 复杂网络研究与复杂现象[J]. 系统辩证学学报，2005，13(4)：8－13.
[3] KRAFFT J，QUATRARO F，SAVIOTTI P P. The knowledge-base evolution in biotechnology：a social network analysis [J]. Economics of innovation and new technology，2011，20(5)：445－475.
[4] CHEN C，IBEKWE-SANJUAN F，HOU J. The structure and dynamics of cocitation clusters：a multiple-perspective cocitation analysis [J]. Journal of the Association for Information Science and Technology，2010，61(7)：1386－1409.
[5] 赵蓉英，王菊. 图书馆学知识图谱分析[J]. 中国图书馆学报，2011，37(2)：40－50.
[6] 安宁，滕广青，白淑春，等. 基于网络 Hub 的领域核心知识涌现研究[J]. 图书情报工作，2017，61(18)：98－106.
[7] 邱均平，张晓培. 基于 CSSCI 的国内知识管理领域作者共被引分析[J]. 情报科学，2011，29(10)：1441－1445.
[8] 刘向，马费成，王晓光. 知识网络的结构及过程模型[J]. 系统工程理论与实践，2013，33(7)：1836－1844.
[9] LIEBOWITZ J. Linking social network analysis with the analytic hi-

erarchy process for knowledge mapping in organizations [J]. Journal of knowledge management, 2005, 9(1): 76 – 86.
[10] KATAKIS I, TSOUMAKAS G, VLAHAVAS I. Multilabel text classification for automated tag suggestion [EB/OL]. [2018 – 07 – 01]. http://lpis.csd.auth.gr/publications/katakis_ecmlpkdd08_challenge.pdf.
[11] GONZALEZ-ALCAIDE G, CASTELLO-COGOLLOS L, NAVARRO-MOLINA C, et al. Library and information science research areas: analysis of journal articles in LISA [J]. Journal of the American Society for Information Science and Technology, 2008, 59 (1): 150 – 154.
[12] ZHANG J, XIE J, HOU W, et al. Mapping the knowledge structure of research on patient adherence: knowledge domain visualization based co-word analysis and social network analysis [J]. Plos one, 2012, 7(4): e34497.
[13] 滕广青, 白淑春, 韩尚轩, 等. 基于无标度与分形理论的层次知识网络原理解析[J]. 图书情报工作, 2017, 61(14): 132 – 140.
[14] SCHVANEVELDT R W, DURSO F T, DEARHOLT D W. Network structures in proximity data [J]. The psychology of learning and motivation, 1989, 24: 249 – 284.
[15] CHEN C, MORRIS S. Visualizing evolving networks: minimum spanning trees versus pathfinder networks [C]//IEEE conference on information visualization. Washington DC: IEEE Computer Society, 2003: 67 – 74.
[16] 韩毅, 童迎, 夏慧. 领域演化结构识别的主路径方法与高被引论文方法对比研究[J]. 图书情报工作, 2013, 57(3): 11 – 16.
[17] AMITH M, CUNNINGHAM R, SAVAS L S, et al. Using Pathfinder networks to discover alignment between expert and consumer conceptual knowledge from online vaccine content [J]. Journal of biomedical informatics, 2017, 74: 33 – 45.
[18] SÁNCHEZ-FRANCO M J, MUÑOZ-EXPÓSITO M, VILLAREJO-RAMOS Á F. A knowledge structures exploration on social network sites [J]. Kybernetes, 2017, 46(5): 818 – 839.
[19] 马瑞敏, 张欣. 基于 Pathfinder 算法的领域知识交流主路径发现研究[J]. 情报学报, 2016, 35(8): 856 – 863.
[20] 张云, 华薇娜, 袁顺波. 利用引文确定领域关键文献的方法探析[J]. 图书情报工作, 2016, 60(1): 66 – 73, 82.
[21] GLEICK J. 信息简史[M]. 高博, 译. 北京: 人民邮电出版社, 2013: 409 – 421.
[22] 李纲, 巴志超. 共词分析过程中的若干问题研究[J]. 中国图书馆学报, 2017, 43(4): 93 – 113.
[23] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine [J]. Computer networks and ISDN systems, 1998, 30(1 – 7): 107 – 117.
[24] FRUCHTERMAN T M J, REINGOLD E M. Graph drawing by force-directed placement [J]. Software: practice and experience, 1991, 21(11): 1129 – 1164.
[25] 滕广青. 基于频度演化的领域知识关联关系涌现[J]. 中国图书馆学报, 2018, 44(3): 79 – 95.
[26] MINTZ B, SCHWARTZ M. The power structure of American business [M]. Chicago: University of Chicago Press, 1987: 224 – 248.

作者贡献说明:
栾宇:数据采集与分析,论文撰写;
滕广青:提出研究思路,设计研究方案,论文撰写与修订;
安宁:数据分析;
韩尚轩:论文修订。

Structural Relationships Extraction of Knowledge Networks Based on Eigen Decomposition

Luan Yu Teng Guangqing An Ning Han Shangxuan
School of Information Science and Technology, Northeast Normal University, Changchun 130117

Abstract: [**Purpose/significance**] The effective identification and extraction of structural relationships in knowledge networks helps to detect the topology of knowledge networks and their evolution patterns from a wide range of data. [**Method/process**] This article proposes a method for extracting structural relationships in knowledge networks based on eigen decomposition of adjacency matrix. Using the real data, the eigen decomposition method and traditional correlation frequency method are compared and analyzed from static structural relationships extraction and dynamic structure evolution, and compared with the pathfinder algorithm. The validity of structural relationships extraction of knowledge networks based on eigen decomposition method is verified. [**Result/conclusion**] The research results show: the eigen decomposition method can identify the main component information in the original knowledge networks, the method can accurately identify the low-frequency correlations that are important to the global topology of the networks, and the extraction method is flexible and free.

Keywords: knowledge network eigen network eigen decomposition structural relationship